복합명사의 역방향 분해 알고리즘

이현민[0] 박혁로
전 남 대 학 교 전 산 학 과
{hyunmini, hyukro}@chonnam.ac.kr

A Reverse Segmentation Algorithm of Compound Nouns

Hyun-Min Lee[0] Hyuk-Ro Park
Dept. of Computer Science, Chonnam National University

요 약

한국어에서 복합명사는 명사간 결합이 자유롭고, 단위명사로 띄어쓰는 것을 원칙으로 하나 붙여써도 무방하다. 따라서, 정보검색분야, 기계번역분야에서 복합명사의 정확한 분해는 시스템의 성능에 많은 영향을 미치게 된다. 본 논문에서는 ETRI의 태깅된 코퍼스로부터 추출한 복합명사를 역방향 분해 알고리즘을 이용하여 단위명사로 분해한다. 분해되지 않은 3119개의 복합명사에 대해 실험한 결과 약 96.6%의 정확도를 얻었다. 또한, 미등록어나 접사에 대한 처리에도 비교적 정확한 결과를 얻을 수 있었다.

1. 서론

자연어 처리 시스템에 있어서 복합명사의 처리는 복잡하고 어렵다. 정보검색분야, 기계번역분야에서 복합명사의 정확한 분해는 시스템의 성능에 커다란 영향을 미친다.

영어의 경우 복합명사의 분해는 비교적 쉬운 반면, 한국어나 독일어의 있어서 복합명사의 분해는 어려운 문제로 인식되고 있다[1]. 그러나, 독일어의 경우도 복합명사를 붙여쓰기는 하지만 단위명사로 쉽게 분해되기 때문에 한국어와 같은 복합명사 분해 문제는 발생하지 않는다. 한국어에서 복합명사는 명사간 결합이 자유롭고, 단위명사로 띄어쓰는 것을 원칙으로 하나 붙여써도 무방하다.

복합명사의 정확한 분해를 위해서는 분해시 발생하는 중의성 문제의 해결과 미등록 단위명사가 포함된 복합명사의 분해 문제, 아울러 분야에 관계없는 정확한 분해 문제가 해결되어야 한다.

본 논문에서는 복합명사의 오른쪽에서부터 왼쪽으로 분해가능한 단위명사로 분해해가는 역방향 분해 알고리즘을 제시한다. 이 분해 알고리즘을 이용하여 ETRI의 태깅된 코퍼스로부터 추출한 복합명사 3119개에 대해 실험한 결과 약 96.6%의 정확도를 얻었다. 이중 444개는 미등록어를 포함한 복합명사로서, 분해 정확도가 77.5%가 되었다.

본 논문의 구성은 다음과 같다. 2장에서는 복합명사와 관련된 기존연구들을 살펴보고 3장에서는 복합명사 분해를 위한 알고리즘을 제시하며 4장에서는 제시한 알고리즘을 이용하여 실험하고 그 결과를 분석하며, 마지막으로 5장에서는 결론을 맺는다.

2. 관련연구

복합명사의 분해에 대한 연구는 크게 통계정보를 이용하는 방법과 구성패턴을 이용한 방법으로 나눌 수 있으며, 두가지를 혼합한 접근 방법도 있다.

[6]에서는 복합명사의 길이에 따른 구성패턴을 파악한 다음, 우선순위가 높은 것부터 적용하는 방식으로 복합명사를 분리하는 방법을 제시하였다. 그러나 복합명사의 길이가 길어질수록 패턴분류에 어려움이 있어, 10음절 이상의 복합명사에 적용하기에는 그리 쉽지 않다.

[4]에서는 복합명사를 먼저 분해패턴에 따라 분해하고, 분해시 발생하는 중의적 분해의 문제를 해결하기 위해서 통계정보와 선호규칙을 이용하는 방법을 제안하고 있다. 통계정보로는 1음절 접사 빈도, 그리고 2음절 또는 3음절 단위명사가 복합명사 내에서 사용된 위치 정보와 빈도 정보를 사용하고, 명사의 개수가 최소로 접속되는 분해패턴을 선호하는 규칙을 적용하였다.

[8]에서는 합성된 상호정보를 이용하여 복합명사를 분해하는 알고리즘을 제안하였다. 합성된 상호정보는 두음절간에 띄어쓴 빈도, 붙여쓴 빈도, 단어의 시작위치에 나타난 빈도, 단어의 끝위치에 나타난 빈도 등 4가지의 상호정보를 합성하여 임계값 이상일 경우 두음절 사이를 띄어쓰고, 이하일 경우에는 붙여쓰는 방법을 제안하고 있다. 이 방법은 복합명사의 음절 길이에 상관없이 적용

될 수 있다.

[7]에서는 원시 코퍼스로부터 단어의 발생 확률 정보를 획득하고, 이 확률 정보를 이용하여 복합명사를 분석하는 비교사학습(unsupervised training)을 이용한 방법을 제안하였다. Hidden Markov Model에 따라 명사사전을 구축하고 구축된 명사사전을 참조하여 복합명사를 분해한다. 만약 분해후보가 둘이상일 때는 색인어로서 가치가 가장 높다고 판단되는 단위명사를 선택하는 방법을 취한다.

[9]에서는 복합명사를 단위명사들로 분해하는 방법으로 네 개의 분해규칙과 두가지 예외규칙을 사용하여 가능한 분해 후보들을 생성하고, 분해 후보들에 대해 가중치를 부여함으로써 최적 후보를 선택하는 알고리즘을 제안하였다. 이 연구에서는 미등록 단위명사가 포함되어 있는 복합명사의 분해뿐만 아니라, 복합명사의 길이에 상관없이 적용되는 방법이 제안되었다.

앞선 연구들의 대부분은 일반적인 복합명사의 분해에 대해서는 비교적 높은 정확도를 보이고 있다. 그러나, 복합명사의 중의적 분해 문제나, 복합명사에 미등록어가 포함되어 있을 경우에는 비록 처리 방안이 제안되어 있기는 하지만, 만족할 만한 수준이 되지는 못한다. 또한, 기존의 복합명사 분해 방법들은 1음절 단위명사에 대한 처리를 고려하지 않았고, 3음절 복합명사의 분해 문제도 배제하고 있다.

3. 복합명사 분해 알고리즘

본 논문에서는 분해되지 않은 복합명사에 대해 사전탐색을 사용하여 적절히 분해할 수 있는 방법을 연구하였다. 적절한 분해를 위해 단위명사 사전과 접사 사전을 이용하였다. 사전탐색은 최장일치되는 단위명사를 우선으로 탐색하도록 처리하였다. 또한, 복합명사의 중심어를 우선 분해하기 위해 분해하는 방향을 오른쪽에서 왼쪽으로 하는 역방향 분해 방법을 적용하였다. 이는 대부분의 복합명사는 대등구조와 종속구조로 구성되며, 종속구조에서의 중심어의 위치는 끝부분에 있는 것을 고려하여 중심어를 먼저 분리해 내기 위해 역방향 분해 알고리즘을 사용하였다.

3.1 단위명사 사전

단위명사 사전은 2음절이상으로 구성된 명사사전을 사용한다. 1음절 명사를 단위명사 사전에 추가하지 않은 것은 한국어에서는 대부분의 1음절어가 단위명사로 사용되기 때문에 복합명사가 너무 짧은 길이의 단위명사로 분해되는 것을 방지하기 위함이다. 예를 들어 '볼가강유역'의 경우, '볼가'가 미등록어이기 때문에 2음절 단위명사로 분해 되지 못하고 '볼'이라는 1음절 명사와 '가'라는 1음절 명사로 인식되어, '볼+가+강+유역'으로 인식되는 오류를 범하게 된다.

3.2 접사 사전

접사 사전은 접두사로 빈번히 사용되는 접사와 접미사로 빈번히 사용되는 접사, 그리고 접두사 및 접미사로 동시에 사용되는 접사를 각각 구분하여 사전을 구성하였다. 그림1과 그림2는 접두사와 접미사의 몇가지 예를 정리한 것이다.

가-	가건물, 가면허, 가문서
고-	고기압, 고소득, 고성능, 고혈압
대-	대가정, 대가족, 대감독, 대강당, 대강령, 대공사
무-	무면허, 무방비, 무사고, 무소속, 무승부, 무시험
비-	비공개, 비공식, 비민주적, 비무장
소-	소개념, 소규모, 소극장, 소문자, 소행성, 소강당
재-	재확인, 재수술, 재시험, 재교육, 재투자, 재작일
전-	전과목, 전국민, 전기간, 전대사, 전세계, 전속력

그림1. 접두사 예제

-가	건축가, 정치가, 소설가, 교육가, 전략가
-계	서무계, 인사계
-고	생산고, 수출고, 판매고
-권	입장권, 상품권. 회수권, 좌석권
-력	생활력, 지도력, 인내력, 경제력, 군사력
-론	관념론, 도덕론, 확률론, 숙명론, 운명론
-률	자동률, 이동률, 희석률, 결합률, 황금률
-소	인쇄소, 발전소, 연구소, 사무소, 강습소, 이발소
-전	개인전, 미술전, 시화전

그림2. 접미사 예제

실제, 한국어에서 사용되는 접사중에는 접미사와 접두사를 모두 사용되는 접사가 많은 비중을 차지하므로, 접두사로 쓸 것인지 아니면 접미사로 사용할지를 결정하는 문제가 발생하게 된다. 앞선 예제에서도 '가', '고'의 경우, 접두사뿐만 아니라 접미사로도 쓰여짐을 알 수 있다. 본 연구에서는 이런 접사에 대해서는 복합명사열에서 접사의 사용위치를 보고 판단하게 하였다. 즉, 접두사와 접미사로 모두 쓰이는 접사일 경우, 복합명사의 첫머리에 오면 접두사로 사용하고, 맨 마지막에 오면 접미사로 사용하며, 가운데에 오면 접미사로 인식하도록 일관성을 부여하였다. 복합명사의 중간에 나타나는 접사중에서 접두사와 접미사로 동시에 쓰일 수 있는 접사를 단순히 접미사로 처리한 이유는 한글의 접사 중에서 접미사의 비중이 접두사보다 높기 때문이다. 예를 들어, 접사 '소'는 접두사와 접미사로 각각 쓰일 수 있는 접사이다. 접사 '소'를 포함하는 복합명사의 분해 예는 다음과 같다.

처음 위치 : 소시민애환 -> 소접두사시민 + 애환
중간 위치 : 분향소설치 -> 분향소접미사 + 설치
마지막 위치 : 직업소개소 -> 직업 + 소개소접미사

3.3 분해 알고리즘

길이가 N인 분리해야 할 복합명사가 들어오면, 오른쪽

에서 왼쪽방향으로 최장일치되는 단위명사를 추출한다. 만약 일치되는 단위명사가 사전에서 발견되지 않으면, 가장 오른쪽 N번째의 1음절을 건너뛴 음절열 skipSyl에 추가하고, 나머지 (N-1)개의 복합명사열을 가지고 다시 사전탐색을 시작한다. 사전탐색에서 최장일치되는 길이가 M인 단위명사를 발견하면, 기존의 건너뛴 음절열 skipSyl에 대해 접사여부를 판별한다. 만약 접두사라면, 이전에 미리 분리된 (i-1)번째 분해열 sp[i-1]에 건너뛴 음절열의 길이 L를 추가하고, 접미사라면 i번째 분해열 sp[i]에 단위명사 길이 M과 건너뛴 음절의 길이 L를 더해서 저장한다. 건너뛴 음절이 접사가 아닐 경우에는 건너뛴 음절의 길이 L을 i번째 분해열 sp[i]에 저장하고, 단위명사의 길이 M을 (i+1)번째 분해열 sp[i+1]에 저장한다. 분해될 길이만큼을 복합명사에서 제거하고 다시 재귀호출 방식으로 복합명사를 분해한다. 이렇게 해서 얻어진 최종 분해열 sp[]는 복합명사의 오른쪽에서 왼쪽 방향으로 분해해야 할 음절 길이 정보를 갖게 된다.

```
char cn[N];  // 분리해야 할 복합명사
int sp[N];   // 역방향으로 저장된 음절 길이 정보
char *skipSyl;       // 건너뛴 음절열
int i;               // 역방향으로 i번째 분리 위치
void segmentCnoun()
{
    char *ptrcn=cn;
    int cutLen;      // cn으로부터 제거할 음절 길이

    cutLen=lookupDictionary(*ptrcn); // 사전탐색
    if (cutLen==0) {
        // 건너뛸 음절의 확장
      skipSyl=strcat(*(ptrcn+strlen(*cn)-1),skipSyl);
      cutLen=1; }
    else
       if (*skipSyl != NULL)  {
          if (isSuffix(*skipSyl))
              cutLen=cutLen+strlen(*skipSyl)
          else if (isPrefix(*skipSyl))
             segPos[i-1]=segPos[i-1]+strlen(*skipSyl)
          else
             segPos[i++]=strlen(*skipSyl);
          segPos[i++]=cutLen
          *skipSyl='₩0'
       }
    *(ptrcn+strlen(cn)-cutLen)='₩0';
    if (cn == NULL)
       if *skipSyl != NULL {
          if (isPrefix(*skipSyl))
             segPos[i-1]=segPos[i-1]+strlen(*skipSyl)
          else
             segPos[i++]=strlen(*skipSyl);
       }
    else
        segmentCnoun();
}
```

그림3. 복합명사 역방향 분해 알고리즘

4. 실험 및 분석

본 연구의 실험을 위해 ETRI의 태깅된 말뭉치로부터 3119개의 복합명사를 추출하였다. 추출한 복합명사를 가지고 역방향 분해 알고리즘을 이용하여 실험한 결과 96.6%의 분해 성공률을 얻었다. 실험한 복합명사의 음절수별 분포 및 정확도는 표1에서 표시한 바와 같고, 그림4는 음절수에 따른 정확도를 도식한 것이다.

음절수	복합명사 개수	분해성공	정확도(%)
3 음절	32 (1.0%)	31	96.8
4 음절	1545 (47.8%)	1540	99.7
5 음절	542 (16.8%)	507	93.5
6 음절	505 (15.6%)	494	97.8
7 음절	215 (6.7%)	196	91.2
8 음절	150 (4.6%)	140	93.3
9 음절	87 (2.7%)	77	88.5
10 음절	52 (1.6%)	49	94.2
11 음절	30 (0.9%)	22	73.3
12 음절	27 (0.8%)	24	88.9
13 음절	14 (0.4%)	12	85.7
14 음절	13 (0.4%)	11	84.6
15 이상	18 (0.6%)	16	88.9
합 계	3119 (100%)	2014	96.6

표1. 복합명사의 음절수별 분포 및 분해 정확도

그림4에서 보는 바와 같이, 음절이 길이가 길어지면 분해의 정확도는 감소해 가는 추세이며, 음절수가 홀수인 경우의 정확도가 짝수의 경우보다 현저히 낮아짐이 발견되었다. 특히, 음절길이가 11인 경우에는 다른 음절길이보다 정확도가 73.3%로 비교적 낮았는데, 이는 '거장니콜라이루빈스타인'이나 '바실리예프스키국방장관'처럼 미등록어로 분리되어야할 명사에, 사전에 등록된 명

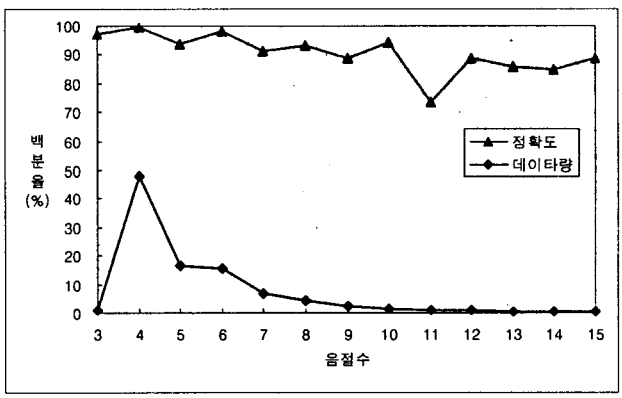

그림4. 음절수에 따른 복합명사의 분해 정확도

사가 포함되어 있어서 분리 가능한 단위명사로 앞뒤를 분리해 버린 경우가 많았기 때문이였다. 이것은 사전에 등록된 단위명사를 포함한 미등록어를 고려하지 않았기 때문이다. 이런 오류는 비단 11음절 복합명사뿐만 아니라 다른 길이의 복합명사에서도 발생하였다. 복합명사의

분해에서 미등록어에 포함된 등록어로 분해하는 오류의 예는 다음과 같다.

알타이미등록어지역 -> 알+타이등록어+지역
투르키스탄미등록어지역 -> 투르+키스등록어+탄+지역
거장니콜라이루빈스타인미등록어 -> 거장+니+콜라등록어+이루빈
+스타등록어인

본 논문에서 사용한 복합명사의 실험자료에는 444개의 미등록어가 포함되어 있다. 이들 미등록어가 포함된 복합명사를 역방향 분해 알고리즘을 이용하여 분해를 수행한 결과 복합명사 344개가 정확히 분해되었고, 분해 성공률은 약 77.5%였다. 미등록어를 포함한 복합명사의 분해 오류는 모두 등록어로만 구성된 복합명사의 분해 성공률 99.6%보다는 매우 낮아서, 미등록어를 포함한 복합명사의 분해 문제가 복합명사 분해 시스템에 중요한 변수로 작용함을 알 수 있었다. 미등록어를 포함한 복합명사의 분해 오류로는 앞서 설명한 미등록어 속에 등록어가 포함되어 발생하는 오류와, '다국적기업직영화미등록어 진출'를 '다국적+기업직접미사 처리+영화+진출'로 분해한 것처럼 접사의 잘못된 결합에 의해서 발생하는 경우가 가장 많았다.

또한, 본 논문에서 제시한 알고리즘의 복합명사의 분해 순서가 역방향이기 때문에 분해 오류가 발생하기도 하였다. 예를 들어 '환자의식'의 경우, 정방향 분해를 했을 경우 '환자+의식'으로 바르게 분해될 수 있으나, 역방향 분해인 까닭에 '자의식'이 '의식'보다 사전에서 먼저 탐색되어 '환+자의식'으로 분해되는 오류가 발생하였다.

그리고, 접사의 처리에서도 잦은 오류가 발생하였다. 접두사 혹은 접미사로 쓰일 수 있는 접사에 대해서 복합명사의 처음에 위치할 때만 접두사로 인식하고, 나머지는 항상 접미사로 인식하도록 하는 일방적인 방법을 사용한 탓에 '민족대화합'의 접사 '대'를 처리하면서, '민족대접미사+화합'으로 분해되는 오류가 발생하였다. 한국어에서 사용되고 있는 대부분의 접사는 접두사와 접미사로서 동시에 사용되는 경우가 대부분인데, 접두사보다는 접미사로서의 비중을 높게 평가하다보니 이런 문제가 발생하였다.

5. 결론

정보검색분야, 기계번역분야 등의 자연어 처리 시스템에서 복합명사를 얼마나 잘 처리하느냐에 따라 시스템의 성능에 커다란 영향을 미친다. 한국어에서 복합명사는 명사간 결합이 자유롭고, 단위명사로 띄어쓰는 것을 원칙으로 하나 붙여써도 무방하기 때문에 복합명사에 대한 처리가 어렵고 복잡하다.

본 논문에서는 ETRI의 태깅된 코퍼스로부터 추출한 복합명사를 역방향 분해 알고리즘을 이용하여 단위명사로 분해한다. 분해후보는 사전탐색을 이용하였으며, 1음절 명사로 인해서 너무 잘게 분해되는 것을 막기 위해, 2음절 이상의 단위명사 사전을 이용하였고, 아울러 접사의 처리를 위해 접사 사전을 사용하였다. 분해되지 않은 3119개의 복합명사에 대해 실험한 결과 약 96.6%의 정확도를 얻었다. 또한, 미등록어를 포함한 복합명사 444개에 대해 77.5%의 분해 성공률을 얻었다.

그러나, 역방향 최장일치 분해를 적용 때문에 발생하는 분해 오류, 접두사 및 접미사로 모두 쓰이는 접사가 복합명사의 가운데에 나타났을 때 일방적으로 접미사로서 인식해서 발생하는 분해 오류, 그리고 긴 미등록어 속에 작은 길이의 등록어가 포함되어 있어서 등록어의 앞뒤에서 분해되 버리는 오류 등은, 제안한 알고리즘의 성능을 저하시키는 중대한 원인으로 작용하였으며, 이러한 문제는 좀더 신중한 검토가 필요할 것으로 사료된다.

6. 참고문헌

[1] T. Pachunke, O. Mertineit, K. Wotheke and R. Schmidt, "Broad Coverage Automatic Morphological Segmentation of German Words," Proceedings of the 14th Conference on Computational Linguistics, pp.1218-1222, 1992.

[2] Bo-Hyun Yun, Ho Lee, Hae-Chang Rim, "Analysis of Korean Compound Nouns Using Statistical Information," Proc. of the 1995 International Conference on Computer Processing of Oriental Languages, pp.76-79, 1995.

[3] H.R. Park, Y.S. Han, K.H.Lee, K.S. Choi, "A Probabilistic Approach to Compound Noun Indexing in Korean Texts," Proceedings of the 16th International Conference on Computational Linguistics, vol.1, pp.514-518, 1996.

[4] 윤보현, 조민정, 임해창, "통계 정보와 선호 규칙을 이용한 한국어 복합명사의 분해", 정보과학회논문지(B), 24권, 8호, pp. 925-928, 1995.

[5] 심광섭, "음절간 상호정보를 이용한 한국어 자동 띄어쓰기", 정보과학회논문지(B), 23권, 9호, pp.991-1000, 1996.

[6] 최재혁, "음절수에 따른 한국어 복합명사 분리 방안", 제8회 한글 및 한국어 정보처리 학술발표논문집, pp.262-267, 1996.

[7] 박혁로, 신중호, "비터비 학습 알고리즘을 이용한 한글 복합명사 분석", 한국정보과학회 학술발표논문집, 1997.

[8] 심광섭, "합성된 상호 정보를 이용한 복합 명사 분리", 정보과학회논문지(B), 24권, 11호, pp.1307-1317, 24권, 11호, 1997.

[9] 강승식, "한국어 복합명사 분해 알고리즘", 정보과학회논문지(B), 25권, 1호, pp172-182, 1998.

[10] 한국전자통신연구원, "전자사전 표제어 선정 지침서", 1999.